DOCUMENT RESUME

ED 319 749 TM 014 833

AUTHOR Xiao, Beiling
TITLE Dichotomous Search Strategies for Computerized Adaptive Testing.
PUB DATE Apr 90
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability Identification; *Adaptive Testing; *Computer Assisted Testing; Equations (Mathematics); Estimation (Mathematics); *Mathematical Models; Maximum Likelihood Statistics; Monte Carlo Methods; Robustness (Statistics); *Scoring; *Search Strategies; Simulation; *Test Construction; Testing Problems; Test Items
IDENTIFIERS Ability Estimates; *Dichotomous Search Strategies; Golden Section Search Strategies; One Parameter Model; Three Parameter Model

ABSTRACT

Dichotomous search strategies (DSSs) for computerized adaptive testing are similar to golden section search strategies (GSSSs). Each middle point of successive search regions is a testing point. After each item is administered, the subject's obtained score is compared with the expected score at successive testing points. If the subject's obtained score does not exceed a confidence interval of an expected score at a testing point, the subject's current ability estimate is assumed to be equal to that of the testing point. Otherwise, the upper or lower half of the search region is discarded and the process is continued until the test-taker's current ability estimate is determined and the next item is selected. Monte Carlo studies with 3,300 subjects in 33 ability levels using one-parameter and three-parameter models to compare the efficiency and accuracy of DSSs, GSSSs, and maximum likelihood estimate strategies (MLESs) indicated that all three measured well in the one-parameter situation, but that DSSs and GSSSs were more accurate and efficient than were MLESs for the three-parameter model. Both DSSs and GSSSs were more robust against guessing and simpler to operate than were MLESs. Twelve graphs illustrate the study. (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED319749

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BEILING XIAO

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."



# Dichotomous Search Strategies for Computerized Adaptive Testing*†

Beiling Xiao
Iowa State University

Paper Presented at the American Educational Research Association 1990 Annual Meeting
April, 16-20, Boston, MA

0* The author is deeply indebted to her major professor Dr. Robert, F. Strahan, minor professor Dr. Yasuo Amemiya, and Dr. Leroy Wolins, Dr. Frederick, G. Brown, Dr. Thomas Andre, for their advice and suggestions on the present research.

0† Correspondence concerning this paper shoul be sent to Beiling Xiao, Department of Psychology, West Lagomarcino Hall, Iowa State University, Ames, IA, 50011-3180.

TM014933

BEST COPY AVAILABLE

## Abstract

Dichotomous search strategies for computerized adaptive testing (DSS) is similar to golden section search strategies for computerized adaptive testing (GSSS) (Xiao, 1989), in which succesive search region shrinks by .5 (by .618 in GSSS). Each middle point of successive search region is a testing point. After each item is administered, Subject's obtained score is compared with expected score at successive testing points. If subject's obtained score does not exceed a confidence interval of a expected score at a testing point, the subject's current ability estimate is assumed equal to that testing point. Otherwise, discard the upper half or the lower half of the search region, and the process continues until the test taker's current ability estimate is determined, upon which the next item is selected to be presented. Monte carlo studies using one-, three- parameter logistic models to compare the efficiency and accuracy of DSS, GSSS, and maximum likelihood estimate strategies (MLES), showed that all three strategies measured very well in one-parameter situation. However, when the three-parameter model was assumed, while guessing existed, both DSS and GSSS were more accurate and more efficient than the MLES. The measurement accuracy of DSS and GSSS were almost the same. Both DSS and GSSS were more robust against guessing, much simpler to operate than the MLES.


* * * * *

Computerized adaptive testing (CAT) uses computer to administer tests, in which the presentation of the next item is adaptive–it depends upon on the test taker's performance on previous items. Each CAT strategy has its own rule to evaluate a test taker's performance, and attempts to provide items that most appropriate for the test taker. CAT methodology offers promise for improving measurement quality in many respects (e.g. , see Weiss, 1982, 1983; Weiss & Kingsbury, 1984; Weiss & Vale, 1987; Lord, 1980; Roid, 1986). CAT can increase measurement precision and efficiency, decrease time for test administration. A general finding is that CAT can reduce test length by an average of 50%, without losing measurement quality. CAT can also control the measurement precision, in which everyone can be measured with a pre-specified degree of precision. Another advantage of CAT is that it can improve standardization of the testing process and test security.

Today, computer technology is developing rapidly. With so many potential advantages, CAT should be also a widely-used method in measurement area. It is not the case. The possible reason might be: earlier CAT strategies that do not based on item response theory (IRT), that are relatively easier to develop, are found much less efficient than IRT-based CAT strategies (see, e.g. , Weiss, 1982; Vale, 1975). CAT strategies that based on IRT models usually need a very large item pool, that meets assumptions of IRT models. To estimate item parameters precisely, large sample size is needed. It is very difficult for individual test makers, such as teachers, to build

such an item pool for a CAT use.

CAT strategies virtually contain three aspects: item selection, ability estimation, and terminate criterion. There are currently two primary IRT-based item scoring methods: maximum likelihood estimate (MLE) (see Lord, 1980, p. 77) and Bayesian scoring approach( Owen, 1969, 1975). There are also two primary IRT-based item selection methods: maximum item information selection and Bayesian item selection. The former chooses an item from the item pool that can provide the maximum value of information at an individual's current ability estimate to administer. The later chooses an item from the item pool that can maximally reduce the posterior variance of an individual's ability estimate when it is administered. There are two primary termination criteria: stop testing when a pre-specified degree of precision is obtained, or when a certain number of items are administered. Maximum information item selection is usually used in conjunction with MLE. While Bayesian item selection is usually used in conjunction with Bayesian scoring. Combining these two item selection methods and ability estimation approaches in a reversed way is also possible. These CAT strategies usually use a pre-specified degree of precision as termination criterion.

Previous studies on maximum information item selection and Bayesian item selection CAT strategies showed that these two kinds of CAT strategies could increase in measurement efficiency, while at the same time achieving equal precision measurements at all levels of the ability continuum (e.g. , Weiss, 1982; Weiss & Kingsbury). MLE procedure try to maximize the information at the current ability estimate level, while Bayesian approach try to minimize the expected posterior variance of the current ability estimate. Information and Bayesian posterior variance are closely related. MLE of ability has asymptotically normal distribution with mean $\theta$, the true ability value, and variance $1/I(\theta)$. It is asymptotically efficient (Birnbaum, 1968, p. 457; Lord, 1980, pp. 70-71). Though they are probably the most sophisticated CAT strategies so far, they have shortcomings. The main disadvantages of MLE strategies include that not all response patterns have MLE solutions or have unique MLE solutions (Samejima, 1973; Lord, 1980, p. 59). When guessing factor exists, in the early stage of a CAT, lower ability subjects can get higher current ability estimates based upon the first few successful item responses. It will take substantial number of items to recover the effect of lucky guessing of the first few items. MLE method also complicated in nature, prohibiting its widespread use. The precision of Bayesian strategies depends upon the accuracy of the pre-ability distribution. When the pre-ability distribution is not accurate, the Bayesian ability estimates are severely biased (Weiss & McBride, 1984).

## Golden Section Search Strategies for CAT

Xiao (1989) proposed golden section search strategies (GSSS) for CAT, that was adapted from golden section search optimization method (see Hua, 1981; Adby & Dempster, 1974; Walsh, 1975, pp. 91-93). In GSSS, test taker's obtained score was compared with expected score of each testing point. Each testing point was located in the middle of each successive golden section search region, and the size of the search region was reduced by the golden section ratio, ($K \approx .618$). If the obtained score was within an confidence interval of an expected score at a testing point, the test taker's current ability estimate was assumed equal to that testing point. Next item was provided according to the current ability estimate. When CAT item number is small ($n = 20$), Xiao found that GSSS measured more accurate in all ability levels but a few extremely high ability levels than MLES did, when the three-parameter logistic model was assumed. GSSS is a hypothesis testing procedure for current ability estimation and item selection. GSSS provided one option for arrangement of testing points. The present research attempts to provide another alternative option.

## Dichotomous Search Strategies for CAT

Dichotomous search strategies (DSS) for CAT is introduced here. Dichotomous search, or dichotoming search, is a procedure for search a point in an interval, in which, at each step, the interval is divided into two halves, one half being then discarded if it can be logically shown that the point could not be in that half. (see, e.g., Hua, 1981; Adby & Dempster, 1974; Gottfried & Weisman, 1973; Parker, 1984) The process of DSS used in this research is as follows:

**testing points.** Testing points are the middle points of successive search regions. The original search region covers all item difficulty levels in the item pool, or all possible ability levels that a test taker's ability estimate would be, whichever is less. The size of a search region is half of the size of the previous one.

**Hypothesis testing and current ability estimation.** After each item is administered, compute the expected score and its variance at each testing point $\theta$. Compute the obtained score of the test taker at each testing point $\theta$. Test scores usually have the following form (see Birnbaum, 1968, Equation 17.7.1):

$$x = \sum_{i=1}^{n} W_i \mu_i \qquad (1)$$

where $W_i$ is the specified numerical weight, and $\mu_i = 1$ when the item response is a key answer, and $\mu_i = 0$ when the item response is not a key answer, n is the number of items.

The mean of $x$ is (see Birnbaum, 1968, Equation 17.7.2):

$$\mu_{x\theta} = \sum_{i=1}^{n} W_i P_i(\theta) \tag{2}$$

The variance of $x$ is (Birnbaum, 1968, Equation 17.7.3):

$$\sigma^2_{x|\theta} = \sum_{i=1}^{n} W_i^2 P_i(\theta) Q_i(\theta) \tag{3}$$

where $P_i(\theta)$ is the item response function of item $i$, and $Q_i(\theta) = 1 - P_i(\theta)$. In three-logistic parameter model, the item response function is (see Lord, 1980, Equation 2-1; Hulin, Drasgow, & Parsons, 1983, Equation 2.4.4):

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta - b_i)]} \tag{4}$$

where $a, b$, and $c$ are item parameters, $D$ is a constant (in this research, $D = 1.702$).

Compare the obtained test score $x$ (at $\theta$) with the expected score $\mu_{x\theta}$ at $\theta$. If $x$ bigger than (or small than) $\mu_{x|\theta}$ plus (or minus) $Z_\alpha \sigma_{x|\theta}$ ($Z_\alpha$ is a prespecified numerical weight), then the test taker's ability is assumed higher (or lower) than the testing point. The lower (or upper) half of the search region is discarded. Hypothesis testing using updated testing point $\theta$ continues, until $x$ is within a confidence interval ($\mu_{x|\theta} \pm Z_\alpha \sigma_{x|\theta}$). The test taker's current ability estimate is assumed to be that testing point $\theta$.

**Test score weights.** The numerical weights of test score in Equations 1 and 2 can be as simple as unit weights (that is 1). If the unit weights are used, the test score is a number-right score. If the weight for each item is chosen by the following formula, the test score can provide ability estimate $\theta$ with optimal or nearly optimal precision (see Birnbaum, 1968, Equation 19.3.3; Lord, 1980, Equation 5-15):

$$W_i = \frac{P_i\prime(\theta)}{P_i(\theta)Q_i(\theta)} \tag{5}$$

where $P_i\prime(\theta)$ is the derivative of $P_i(\theta)$ with respect to $\theta$. The weights are called locally best weights by Birnbaum, and optimal scoring weights by Lord. If we substitute the optimal scoring weights into the following equation:

$$\sum_{i=1}^{n} W_i \mu_i = \sum_{i=1}^{n} W_i P_i(\theta) \tag{6}$$

it becomes

$$\sum_{i=1}^{N} \frac{P_i\prime(\theta)}{P_i(\theta)Q_i(\theta)} \mu_i = \sum_{i=1}^{n} \frac{P_i\prime(\theta)}{Q_i(\theta)} \tag{7}$$

Equation 7 is a maximum likelihood equation. A test taker's MLE of ability $\theta$ can be obtained by solving Equation 7, if the response pattern of the person is known. The left side of Equation 6 is the test taker's obtained score, the right side of Equation 6 is the expected score at ability $\theta$. If optimal weights are used in DSS, The current ability estimate is within a confidence interval of $\theta$, that is, the confidence interval of MLE.

**Lower than chance score adjustment.** Equation 4 has a pseudo-guessing parameter $c_i$. Whenever a test taker's obtained score $x$ is less than a portion of score assumed from random guessing

$$x_c = \sum_{i=1}^{n} W_i c_i \tag{8}$$

The obtained score $x$ is adjusted into $x_c$. This adjustment is only applied during item selection process. In final ability estimation, no such adjustment is made.

**Item selection.** Quasi-match $m_i$ to $\hat{\theta}$ item selection is used whenever a current ability estimate is made. When there is no guessing, item whose difficulty is equal to the current ability estimate $\hat{\theta}$ can maximize the item information at $\hat{\theta}$. While in three-parameter logistic model, when guessing occurs, item whose difficulty slightly lower than $\hat{\theta}$ can maximize the item information at $\hat{\theta}$ (see Lord, 1980, pp. 172-173). In DSS, a constant $d$ is calculated for each item pool:

$$d = \frac{1}{Da} \ln \frac{1 + \sqrt{1 + 8c_i}}{2} \tag{9}$$

where $a$ and $c$ are the means of $a_i$ and $c_i$ of the whole item pool, respectively. Clearly, $d = 0$ when $c_i$ of each item is 0. That is the case in one-, or two-parameter logistic model.

define $m_i = b_i + d$. After each current ability is determined, an item is chosen to administer, whose $m_i$ value is closest to $\hat{\theta}$. If an item to be selected is already been administered, an item which is next easier (or next more difficult) to it, is chosen from the item pool, depending upon whether the obtained score $x$ is smaller (or greater) than the expected score at $\hat{\theta}$.

**Test termination criterion.** Both two primary termination criteria discussed previously could be used in DSS CAT: a pre-specified degree of precision, or certain number of items. The index used in the former criterion could be test information (see Lord, 1980, pp. 65-80) at current ability estimate, or response pattern information at current ability estimate (see Samejima, 1973, 1983, pp. 159-165).

**Final ability estimation.** MLE scoring method could be used in DSS to obtain the final score. A lot of methods could apply to solve the MLE solution. one of them is modified Newton-Raphson method (Hamming, 1973, pp. 68-72). Another simplest method is to apply dichotomous search to solve the MLE of ability.

The above description about DSS showed the robust nature of it, and its potential to measure effectively. The present monte carlo research try to testify the efficiency and accuracy of DSS. DSS is similar to GSSS in nature, only differ in the ratios of their successive search regions. Monte carlo studies are conducted to compare efficiency and accuracy of three CAT strategies: DSS, GSSS, and MLES. It is expected that DSS, as well as GSSS, will provide more precision ability estimate than MLES does.

## Method

### Design

The precision of ability estimates of three CAT strategies: DSS, GSSS, and MLES, in four computer simulation studies, was compared. Each CAT contained 20 items. Three estimated precision indices–response pattern information, bias, and mean squared error, were compared at each chosen ability levels among three strategies within each study. One-parameter logistic model was assumed in Study 1. In Studies 2, 3, and 4, three-parameter logistic model were assumed, respectively. All four studies were done by ZENITH micro-computers. All programs were programmed by the author using GWBASIC language, and were compiled into machine language.

Samejima (1973, pp. 226-228) explained that item response information function is an accuracy index for each response category in estimation $\theta$. Item information function is the expectation of the item response information function. Response pattern information function is the sum of each item response information function. It is a direct measure of accuracy for each individual's response pattern. The formula to calculate response pattern information for response pattern $V$, in three-parameter logistic model is

$$I_V(\theta) = -D^2 \sum_{i=1}^{n} \frac{a_i^2 c_i \exp[-Da_i(\theta - b_i)]}{\{1 + c_i \exp[-Da_i(\theta - b_i)]\}^2} \mu_i + D^2 \sum_{i=1}^{n} \frac{a_i^2 \exp[-Da_i(\theta - b_i)]}{\{1 + \exp[-Da_i(\theta - b_i)]\}^2} \tag{10}$$

in one-parameter logistic model is

$$I_V = D^2 \sum_{i=1}^{n} \frac{\exp[-D(\theta - b_i)]}{\{1 + \exp[-D(\theta - b_i)]\}^2} \tag{11}$$

The bias is defined as

$$B(\theta) = \frac{1}{N} \sum_{i=1}^{n} (\hat{\theta}_i - \theta) \tag{12}$$

The mean squared error (MSE) is

$$MSE(\theta) = \frac{1}{N}\sum_{i=1}^{n}(\hat{\theta}_i - \theta)^2 \qquad (13)$$

**Subject**

An subject was a computer generated test taker with a true ability value $\theta$. In each of the four studies, there would be 3300 subjects in each of three CAT strategies, with 100 as a group at each of 33 ability levels equally spaced in [-3.2, 3.2], in interval of 0.2.

**Item Pool and Item Responses**

There was a common item pool for the three strategies in each study. Each item pool had 181 items, whose difficulties were equally spaced from -3.6 to 3.6, in interval 0.04. Item discrimination parameters and item pseudo-guessing parameters in each item pool were constants. In Study 1, $a_i = 1$, $c_i = 0$; in Study 2, $a_i = 1.25$, $c_i = .2$; in Study 3, $a_i = 1.25$, $c_i = .25$; in Study 4, $a_i = 1.25$, $c_i = .333$.

When an item was provided to a subject whose ability was $\theta$, a random number $g$ generated from a distribution uniform in interval (0, 1), was compared with the item response function $P_i(\theta)$ (see Equation 4, one-, or two-parameter logistic model is a special case of the three-parameter logistic model). Item response $\mu_i = 1$ when $g$ was smaller than or equal to $P_i(\theta)$; otherwise, $\mu_i = 0$.

**CAT Strategies**

In all three CAT strategies–DSS, GSSS, and MLES, all subjects were assumed a pre-ability estimate $\theta = 0$. Modified Newton-Raphson method (Hamming, 1973, pp. 68-72) was used in final ability estimation to solve MLE of ability, after 20 items was administered. Iteration process continued until the difference between two successive estimated $\theta$ values was less than 0.001. If a subject's item responses were all 1s or all 0s, or a subject's MLE solution was less than -10, or greater than 10, it would be excluded from the results. Each excluded case was replaced by a new subject with the same ability level, to maintain 100 subjects in each group.

**DSS** In DSS, the original search region was the same for all subjects [-3.6, 3.6]. The size of successive search region reduced by ratio .5. Items were chosen according to the quasi-match $m_i$ to $\hat{\theta}$ rule. After each item was administered, hypothesis tests were conducted, from the original search region on, until a testing point was found, at which the confidence interval of expected score $\mu_{x|\theta}$, covered the obtained score. Predefined $Z_\alpha = 1$, therefore, the confidence interval is $(\mu_{x|\theta} \pm \sigma_{x|\theta})$. In situation where the search region was reduced to less than 0.08, the subject's current ability is assumed to be equal to that last testing point, regardless the hypothesis

testing results.

**GSSS** GSSS was the same as DSS, excerpt for the ratio of sizes of successive search regions. In GSSS, search region reduced by ratio $\frac{\sqrt{5}-1}{2} \approx .618$. There were two cutting points in each search region. If the search region is [0, 1], then the two cutting points were $\frac{\sqrt{5}-1}{2} \approx .618$, $1 - \frac{\sqrt{5}-1}{2} \approx .382$. Similar to DSS, each testing point was at the middle of each search region. Hypothesis tests determined whether the upper (or lower) portion of the search region, from the upper (or lower) cutting point, needed to be discarded; or the current testing point was the current ability estimate.

**MLES** MLES used modified Newton-Raphson method to solve MLE after each item was administered. Maximum information item selection method was use in item selection. During item selection process, If a subject's item responses were all 1s (or all 0s), or subject's current ability estimate was bigger than 3.8 (or less than -3.6), the subject's current ability estimate was assigned to be 3.8 (or -3.6).

## Results

Results of present research were as expected. When guessing existed, DSS, as well as GSSS, could provide more accurate and more efficient ability estimates than MLES did along the whole ability continuum, except for a few extremely high ability levels. DSS and GSSS had similar measurement accuracy. DSS measure slightly better than GSSS in the two ends of ability continuum than GSSS; GSSS provided slightly more accurate estimates than DSS in the middle range of ability than DSS did.

### Response Pattern Information

Calculated the averaged response pattern information of 100 subjects at each 33 levels of ability for each CAT strategy, in each study, then averaged the 33 means. The overall mean of response pattern information for each CAT strategy in each study for DSS, GSSS, and MLES, were 11.649, 11.667, and 11.565 in Study 1; 10.665, 10.639, and 10.182 in Study 2; 9.332, 9.303, 8.880 in Study 3; 7.602, 7.586, and 6.919 in Study 4, respectively.

---

Insert Figure 1 about here

---

Figure 1 showed the response pattern information curves of DSS, GSSS, and MLES in Study 1. The average height of the three curves were about the same. In the middle range of ability levels (about -1.8 to 2) response pattern information curve of GSSS was slightly higher than that of DSS's, both information curves of DSS and GSSS were higher than that of MLES; At the two ends of the ability continuum,

MLES have the highest response pattern information curves among the three CAT strategies, while the curve of DSS was slightly higher than the curve of GSSS.

---

Insert Figures 2, 3, and 4 about here.

---

Figures 2, 3, and 4 showed response pattern information curves in Studies 2, 3, and 4, respectively. The response pattern information curves of DSS and GSSS were almost overlaped in each study. The average heights of MLES's were lower than those of DSS or GSSS. In study 1, or 2, or 3, The curves of DSS or of GSSS were higher in all ability levels, except for a few extremely higher levels of ability and a small range of extremely lower ability levels. The height of response pattern information curve of GSSS was slightly higher than those of DSS in the middle range of ability in each of these studies; while in the two ends, DSS had slightly higher response information curves.

---

Insert Figure 5 about here

---

**Bias**

Bias at each of 33 ability levels for each of three strategies in each of four studies were computed. Bias in the four studies were not big for all three strategies, except for bias of MLES at a few lower ability levels when guessing existed. Figure 5 showed the bias in Study 1. The bias were very small in all three strategies. Most of the bias were not exceed interval (-0.05, 0.05).

---

Insert Figures 6, 7, 8 about here

---

Bias in Study 2 and 3 were also not big for three strategies, except for a few relatively big bias in extremely lower ability levels for MLES. From Figures 6 and 7, It could be seen that Most bias for the three strategies were in interval (0, 0.1) in Study 2; in interval (-0.03,0.1) in Study 3. Figure 8 showed the bias in Study 4 for the three strategies, in which the pseudo-guessing parameter $c_i$ was as big as .333, bias of DSS and of GSSS were relatively smaller than that of MLES, most bias of DSS and of GSSS were in interval (-0.02, 0.1); while most bias of MLES were in interval (0.05, 0.2). A few relatively big bias occurred in the lower ability levels in MLES, some exceeded 0.3.

---

Insert Figure 9 about here

---

### MSE

MSEs at each level of ability for each CAT strategy in each study were calculated. Computed mean MSEs of all 33 ability levels for each strategy in each study. The mean MSEs for DSS, GSSS, and MLES, were 0.082, 0.087, 0.086 in Study 1; 0.116, 0.116, 0.172 in Study 2; 0.129, 0.127, 0.217 in Study 3; 0.189, 0.166, 0.386 in Study 4, respectively. Figure 9 showed MSE curves in Study 1. Three MSE curves were overlaped. The MSEs at all ability levels were very small for all three strategies. The curves were very flat.

---

Insert Figures 10, 11, and 12 about here

---

Bigger MSEs were found in MLES than those in DSS and GSSS in Studies 2, 3, and 4, when guessing existed. The more pseudo-guessing value $c_i$, the more such difference. From figures 10, 11, and 12, it could be seen that MSEs of DSS and GSSS were almost the same. MSEs of MLES were bigger than those of DSS and GSSS, except for a few very higher ability levels. Significant differences of MSEs between DSS (or GSSS) and MLES were found almost along the whole ability continuum in Studies 3 and 4; in Study 2, such difference occurred in the lower and middle range of ability.

### Deleted Cases

As mentioned previously, subjects whose ability estimate bigger than 10 or smaller than -10, or whose responses to all 20 items were all 1s or all 0s, were excluded from the research. Other subjects with the same ability levels had replaced them. In study 1, no subject was replaced; the number of replaced subjects for DSS, GSSS, and MLES, were 0, 5, 4 in Study 2; 1, 3, 4 in Study 3; 5, 11, 4 in Study 4, respectively. The majority replacement occurred in extremely lower ability subjects.

## Discussion

Results of the present research showed robust and precise nature of DSS and GSSS –they were quite robust against guessing; robust with respect to relative inaccuracy of item parameter estimation; the potential to provide accurate and efficient ability estimate while in the same time simplify the CAT process; the potential to enable CAT to be more popular, especially in educational settings.

There are many robust features of DSS. In DSS, successive testing points are defined sequentially, upon which test taker's current ability is determined. In determining each current ability estimate, a confidence interval of expected score at a testing point is used, to evaluate whether the obtained score is equal to the expected score at that testing point. If a very very small confidence interval is used, and the

optimal weights are used in hypothesis testing, the testing point, determined by DSS as a current ability estimate, will be very very close to MLE of ability at that stage. Since the last very very small confidence interval contains the MLE. We can see that DSS estimate during item selection process is more conservative than MLE does, the current ability estimate tends to remain in the previous estimate position. Test taker who makes a few lucky guessing at the earlier stage of CAT will not get very high current ability estimate. There is also an adjustment for lower than chance scores, that prevents test taker get very low ability estimate during the earlier stage of CAT. After each item is administered, hypothesis testing is conducted from the original search region on, until a current estimate is determined. The original search region, is assumed to cover the range of possible ability estimate of that test taker, or the difficulty range of the item pool. In that manner, test taker who misses a few items by mistake, or who answered successfully by lucky guessing, in the earlier stage of CAT, will have substantial chance to achieve an ability estimate that would appropriate to him/her.

DSS uses quasi-match $m_1$ to $\hat{\theta}_i$ item selection to provide items for testing. An item that has $m_i$ value most closest to $\hat{\theta}_i$ will be chosen, after each current ability is determined. The difference between $m_i$ and the item's difficulty $b_i$ is a constant for each item pool, If an item being selected is already administered for that particular test taker, an item that is next easier (or next more difficult) to it, will be selected, depending upon whether the obtained score is smaller (or greater) than the expected score at a testing point. Since an item pool used in practice contains finite number of items. If items are ordered by their difficulties, there usually is a difficulty gap between any two successive items. A most efficient test for a particular individual, chosen from an item pool, will contain items whose difficulties are around the individual's true ability value. The difficulties of items in the test will have a relatively wide range. For example, the optimal test for the present study will cover difficulty width 0.76. The DSS's current ability estimate is not necessary the expected score at the testing point, but in its confidence interval. By choosing items in right direction (that is, towards the expected value) DSS is very efficient to provide appropriate items for test taker. Most items chosen for an individual have successive difficulties. By an MLE final ability estimation procedure, the DSS can get optimal or nearly optimal ability estimate.

In this research, item parameters were assumed estimated without error. In practice, it was not true. A confidence interval in each hypothesis testing enables DSS to be robust with respect to relatively poor item parameter estimation. Optimal weights were used in the present studies, it enabled DSS to provide very accurate estimate along the ability continuum. In DSS, If unit weight was used, the obtained score (number-right) would be towards the number-right expected score at a current

ability estimate. That expected score was the sum of each $P_i(\theta)$; the variance of the number-right score was simply the sum of products of each $P_i(\theta)$ and $Q_i(\theta)$. Though it might lose some efficiency, using unit weight in DSS could always provide an estimate that would be towards the true ability.

MLE has some optimal asymptotic properties. "Nothing very much can be said about the MLE small sample performance (Norden, 1973, p. 47)". A current ability estimate based on very small samples might be misleading. DSS, as well as GSSS, used more conservative current ability estimates than MLES did, and got equal precision results compared to MLES in non-guessing situation; achieved more accurate results whenever guessing occurred. The response pattern information curves of MLES were higher in both ends of ability continuum, but MSEs obtained by MLES were always greater than those of DSS and GSSS in lower ability range in studies 2, 3, and 4. The reciprocal of MSE could be an estimate of test information (see Lord, 1980, pp. 89) in this research, since all three strategies used MLE to solve final ability estimate, and MLE was unbiased asymptotically. The author also recorded average test information for each group, which was almost the same as response pattern information for each group in Studies 2, 3, and 4; (in study 1, in which $c_i = 0$, the two information indices were exactly the same). The controversy of response pattern information and MSE results might be due to the fact that the average response pattern information of each group would not be affected very much by a few test taker's unusual response patterns (such as lower ability ones answered difficult items correctly), while MSE might be much greater if a few ability estimates were far away from their true values.

DSS, as well as GSSS, established a systematic way to control the CAT process statistically. The present studies were only conducted in simulated situations. More research is needed to apply DSS and GSSS in real test situations.

## References


Adby, P. R., & Dempster, M. A. H. (1974). *Introduction to optimization methods.* (pp. 18-41). Landon: Chapman and Hall.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical Theories of mental test scores* (Part 5, pp. 397-479). Reading, MA: Addison-Wesley.

Gottfried, B. S. & Weisman, J. (1973). *Introduction to optimization theory* (pp. 66-135). Englewood Cliffs, NJ: Prentice-Hall.

Hamming, R. W. (1973). *Numerical methods for scientists and engineers* (2nd). New York: McGraw-Hull.

Hua, Luogen (1981). 优选法 (*Optimization methods*) (pp. 3-32). Beijing: Science Press.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Norden, R. H. (1973). A survey of maximum likelihood estimation, Part 2. *International Statistical Review, 41,* 39-58.

Parker, S. P. (Ed.). (1984). *McGraw-Hill Dictionary of Science and Technical Terms* (3rd.). New York: McGrqw-Hill.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Roid, G. H. (1986). Computer technology in testing. In B. S. Plake & J. C. Witt (Eds.). *The future of testing* ( pp. 29-69). Hillsdale, HJ: Lawrence Erlbaum.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38,* 221-233.

Samejima, F. (1983). Some methods and approaches of estimating the operating characteristics of discrete item responses. In F. M. Lord (Ed.) *Principles of psychological measurement* (pp. 159-182). Hillsdale, NJ: Lawrence Erlbaum.

Vale, C. D. (1975). Problem: Strategies of branching through an item pool. In D. J. Weiss (Ed.). *Computerized adaptive trait measurement: Problems and prospects* (research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Appllied Psychological Measurement, 6,* 473-492.

Weiss, D. J. (1983). Introduction. In D. J. Weiss (Ed. ). *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 1-8). New York: Academic.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of compuerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361-375.

Weiss, D. J., & McBride, J. R.(1984). Bias and information of Bayesian adaptive Testing. *Applied Psychological Measurement , 8,* 273-285.

Weiss, D. J., & Vale, C. D. (1987). Adaptive testing. *Applied Psychology: An International Review, 36,* 249-262.

Xiao, Beiling (1989, April). *Golden section search strategies for computerized adaptive testing.* Paper presented at the Fifth International Objective Measurement Workshop, Berkeley, CA.

Walsh, G. R. (1975). *Methods of optimization.* London: John Wiley and Sons.

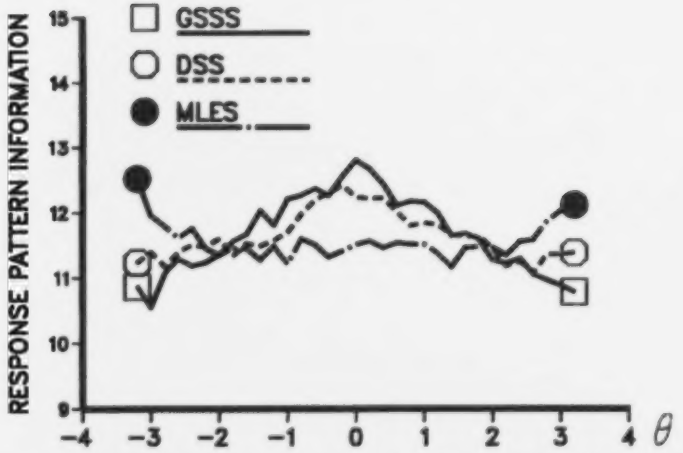


Figure 1. Response Pattern Information Curves of GSSS, DSS, and MLES in Study 1 ($a = 1.00$, $c = 0$)

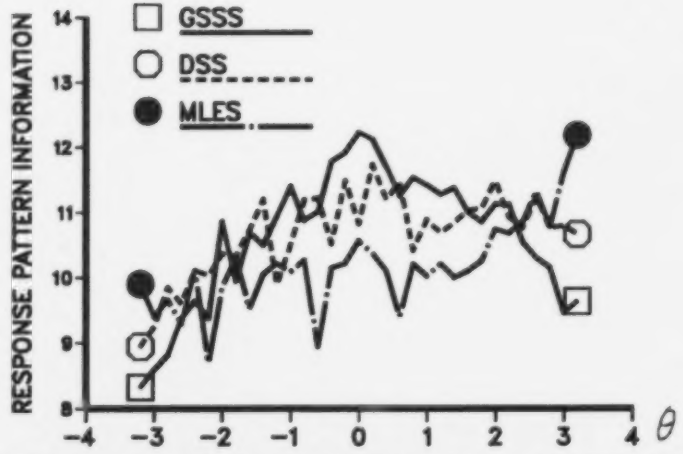


Figure 2. Response Pattern Information Curves of GSSS, DSS, and MLES in Study 2 ($a = 1.25$, $c = .20$)

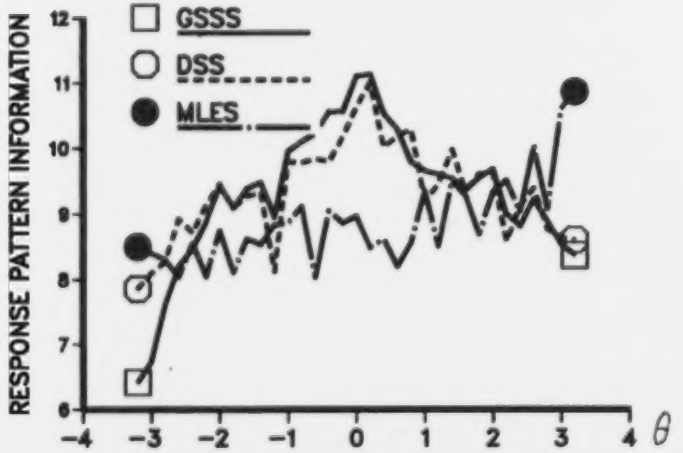


Figure 3. Response Pattern Information Curves of GSSS, DSS, and MLES in Study 3 ($a = 1.25$, $c = .25$)

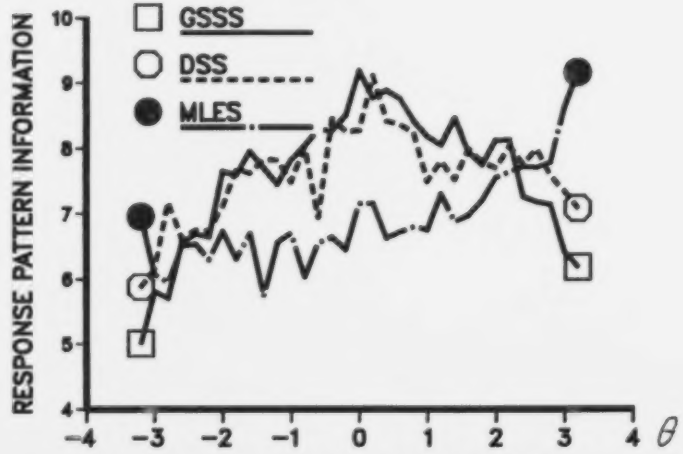


Figure 4. Response Pattern Information Curves of GSSS, DSS, and MLES in Study 4 ($a = 1.25$, $c = .333$)

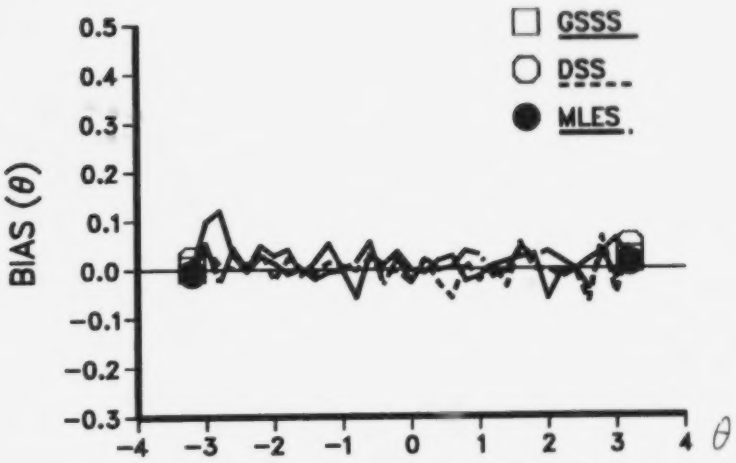


Figure 5. Bias of GSSS, DSS, and MLES in Study 1 ($a = 1.00$, $c = 0$)

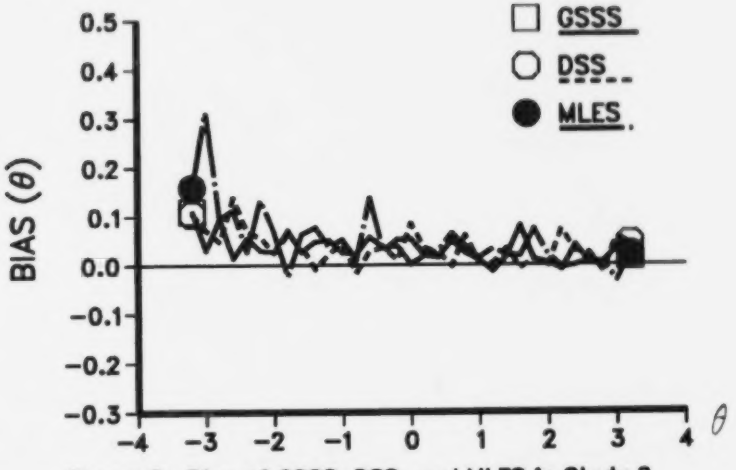


Figure 6. Bias of GSSS, DSS, and MLES in Study 2 ($a = 1.25$, $c = .20$)

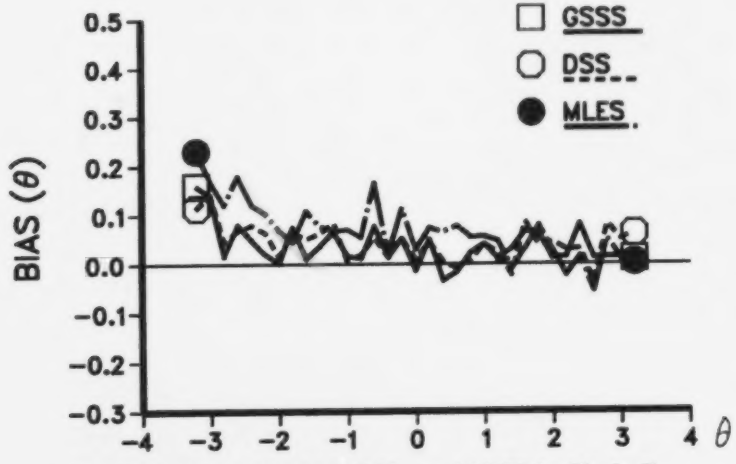


Figure 7. Bias of GSSS, DSS, and MLES in Study 3 ($a = 1.25$, $c = .25$)

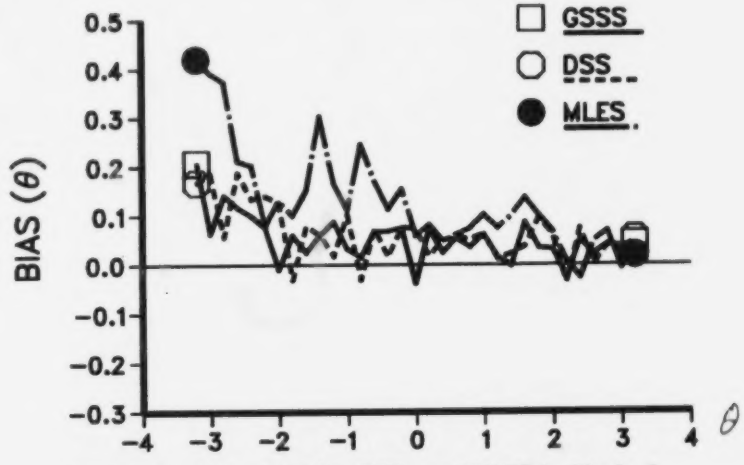


Figure 8. Bias of GSSS, DSS, and MLES in Study 4 ($a = 1.25$, $c = .333$)

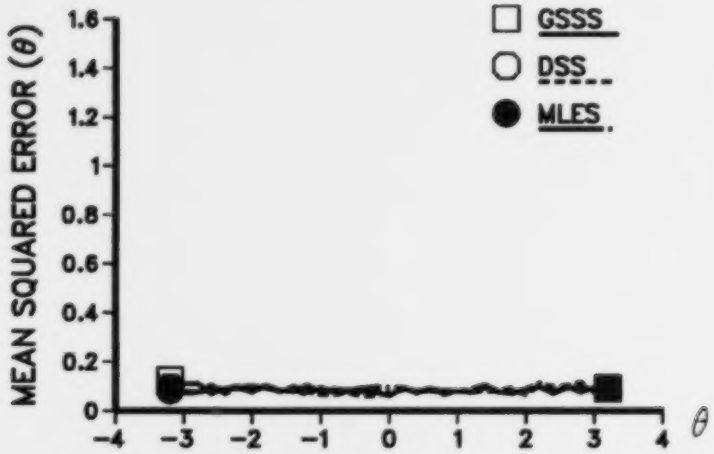


Figure 9. Mean Squared Errors (MSE) of GSSS, DSS, and MLES In Study 1 ($a = 1.00$, $c = 0$)

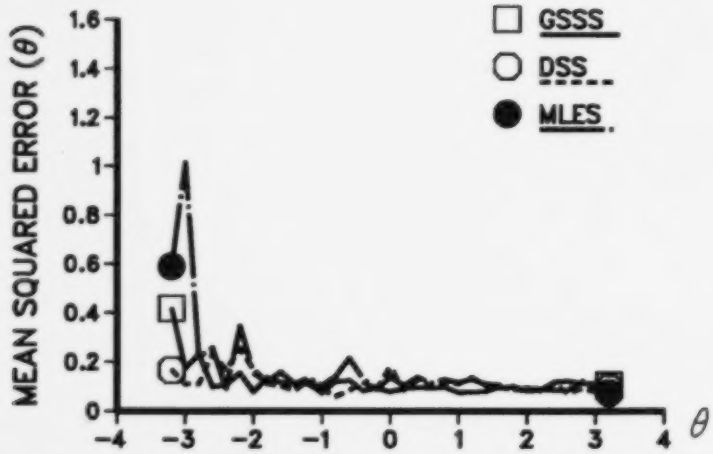


Figure 10. Mean Squared Errors (MSE) of GSSS, DSS, and MLES In Study 2 ($a = 1.25$, $c = .20$)

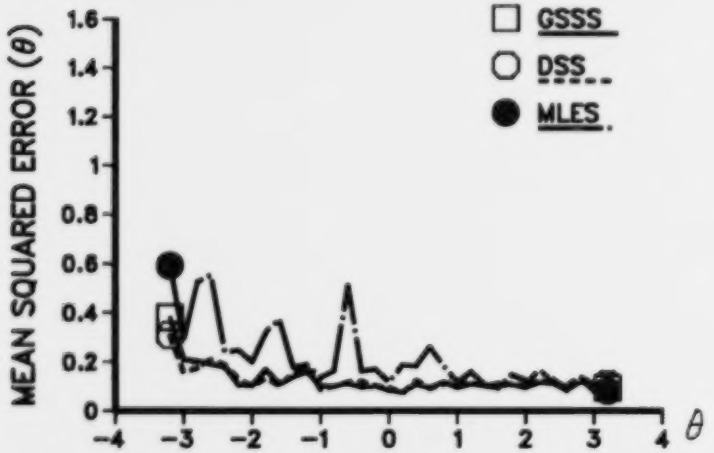


Figure 11. Mean Squared Errors (MSE) of GSSS, DSS, and MLES In Study 3 ($a = 1.25$, $c = .25$)

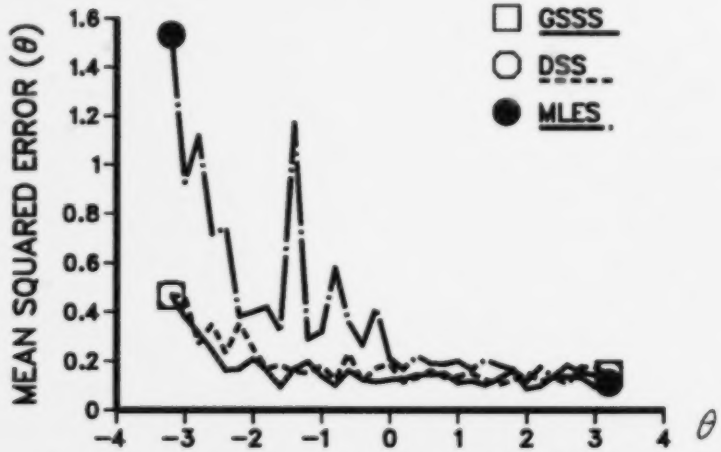


Figure 12. Mean Squared Errors (MSE) of GSSS, DSS, and MLES In Study 4 ($a = 1.25$, $c = .333$)